DOCUMENT RESUME

ED 175 882 TM 008 449

AUTHOR Watkins, John M.: And Others
TITLE The Generalizability of the Matching Familiar Figures Test.
PUB DATE [Mar 78]
NOTE 26p.: Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Age Differences: *Cognitive Development: *Cognitive Tests: *Conceptual Tempo: Elementary Education: Males: Reaction Time: Scores: *Test Reliability
IDENTIFIERS *Generalizability Theory: *Matching Familiar Figures Test (Kagan): Matching to Sample Procedure: Test Length

ABSTRACT

Generalizability theory was applied to the Matching Familiar Figures Test (MFF), an instrument commonly employed to assess reflection-impulsivity in children, in order to analyze the dependability of the MFF at four grade levels: second, third, fourth, and fifth. The MFF was individually administered to 114 boys. A completely crossed, two-facet random model design was used to provide a multidimensional framework for examining the dependability of the MFF. Coefficients of generalizability were derived from this design for the MFF error and latency scores at each grade level. Results showed that the MFF latency score was a more dependable measure than the MFF error score. In addition, the number of testing occasions made a more significant contribution to the generalizability of the MFF than the number of items. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED175882

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Generalizability of the Matching Familiar Figures Test

John M. Watkins
Los Angeles Unified School District
Howard B. Lee
University of California, Los Angeles
Oded Erlich
Tel Aviv University

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
John M. Watkins
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the 1978 Annual Convention of the American Educational Research Association, Toronto, Canada.
Session 7.04

TM008 449

Abstract

Generalizability theory was applied to the Matching Familiar Figures Test (MFF) to illustrate its applicability to an instrument commonly employed to assess reflection-impulsivity in children and to analyze the dependability of the MFF at four grade levels: second, third, fourth, and fifth. A completely crossed, two-facet random model design was used to provide a multidimensional framework for examining the dependability of the MFF. Coefficients of generalizability were derived from this design for the MFF error and latency scores at each grade level. Results showed that the MFF latency score was a more dependable measure than the MFF error score. In addition, the number of testing occasions made a more significant contribution to the generalizability of the MFF than the number of items.

The Generalizability of the Matching Familiar Figures Test

The Matching Familiar Figures Test (MFF) (Kagan, Rosman, Day, Albert, & Phillips, 1964) has received considerable attention over the past decade as a measure of reflection-impulsivity. Despite the attention accorded the MFF in the literature, only recently has consideration been given to the formal psychometric properties of the MFF (Ault, Mitchell, & Hartmann, 1976; Egeland & Weinberg, 1976).

The present investigation used generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to provide a multidimensional view of the dependability of the MFF with second through fifth grade boys. The present investigation has three main purposes: (1) to illustrate the application of generalizability theory to an instrument frequently employed in the child development literature, (2) to analyze the dependability of the MFF at four different grade levels, and (3) to examine the implications of the MFF's generalizability for enhancing our understanding of past studies which have used the MFF and to provide guidelines for increasing the dependability of the MFF in

future research.

The Matching Familiar Figures Test

The MFF was developed by Kagan et al. (1964) to measure the reflection-impulsivity dimension, which is "concerned with the degree to which the subject reflects on the validity of his solution hypotheses in problems that contain response uncertainty" (Kagan & Kogan, 1970, p. 1309). The MFF is a match-to-sample task containing a picture of a familiar object, below which are arranged six similar variants, one of which is identical to the standard. The child is instructed to select the one variant that is identical to the standard. Two dependent variables, latency to first response and total number of errors, are recorded. A moderate negative correlation generally obtains between latency and error scores (Kagan, 1965; Kagan & Messer, 1975).

Recently, several investigators have examined the validity (Block, Block, & Harrington, 1974; Kagan & Messer, 1975; Bentler & McLain, 1976; Zelniker, Bentler, & Renan, 1977) and reliability (Ault et al., 1976; Egeland & Weinberg, 1976) of the MFF. All of these studies have approached the validity and reliability of the MFF from the perspective of classical test theory.

The most extensive examination of MFF reliability has been presented by Egeland and Weinberg (1976). They evaluated the short-term reliability of three different forms of the MFF with kindergarten, second, and fifth grade boys and girls. Test-retest correlational analyses for latency scores were based on logarithmically transformed scores. Product-moment correlations for latency scores separated by a one week inter-test interval ranged from .41 (kindergarten girls, Form K) to .78 (fifth grade girls, Form F). Test-retest correlations for error scores ranged from .27 (fifth grade girls, Form F) to .77 (second grade girls, Form F). Egeland and Weinberg concluded that "test-retest reliability for the MFFT is fair at the second and fifth grades and poor at kindergarten" (p. 487).

The Egeland and Weinberg (1976) data were not consistent in showing differences in reliability between latency and error. For example, latency was less reliable than error at kindergarten, but at the fifth grade error was less reliable than latency. Ault et al. (1976), however, report a consistent pattern of low test-retest and internal consistency reliability for MFF error scores.

Generalizability Theory

Cronbach et al. (1972), expanding on classical test theory, have developed a multifacet method of analyzing sources of error for behavioral measurements. Like classical test theory, generalizability theory addresses the question of how well one can generalize from a sample of behavior to the characteristic behavior one is attempting to measure. The sample of behavior, or observed score, is taken to be more or less representative of an ideal which is termed the universe score. In generalizability theory, the true score is conceived of as a universe score which is defined as the mean score an individual would obtain if observed over the entire universe. Generalizability theory thus defines the accuracy of a behavioral measurement in terms of how well the observed score generalizes to the universe of admissible observations. This conceptualization is essentially an expansion of the classical notion of reliability: "The question of 'reliability' thus resolves into a question of accuracy of generalization, or generalizability" (Cronbach et al., 1972, p. 15).

The various conditions which might constitute sources of error are referred to as facets. The levels of each

facet are termed conditions of the facet. The cartesian product of the facets defines the universe of admissible observations. Thus, any given observation is defined as a combination of conditions that define the separate facets.

Generalizability theory recognizes two principal types of studies: the generalizability (G) study and the decision (D) study. For a G study, the investigator develops a design based, within practical limits, on all possible facets which may influence the generalizability of the measurement. The data then provide the basis for estimating components of variance which in turn are used to calculate the coefficient of generalizability. Based on information provided in the G study, a D study can then be developed to provide the basis for substantive decisions.

Generalizability theory has several advantages over the more commonly used methods of evaluating reliability. In classical test theory, a person's observed score, X, is said to be a linear combination of his true score, T, and random error, E: $X = T + E$. The observed score variance is held to be equal to the true score variance plus error variance. The reliability of a measurement is equal to the ratio of true score variance to observed score variance. However, the reliability coefficient can be

estimated in a number of different ways (e.g. test-retest, KR-20). The crucial feature of these classical approaches to reliability is that each method defines the error that enters into an observed score in a different way.

In contrast, generalizability theory has the advantage of isolating several sources of variance simultaneously, thus providing the basis for relative decisions. A more comprehensive picture of the combined effects of all relevant sources of error is obtained and the relative potency of different sources of variation becomes known. In addition, generalizability theory takes into consideration the variance contributed by interactions.

The central concept underlying the systematic partitioning of error variance in generalizability theory lies in the recognition that scores can generalize over different universes. Cronbach et al. (1972) illustrate the advantage of this procedure over classical test theory:

> Any observation...fits within a variety of universes. "The universe score is estimated to be 75" is without meaning until we answer the question, "Which universe?" This ambiguity is concealed in the statement "The estimated true score is 75," for no one thinks to inquire, "Which truth?" (p. 18-19).

The Two-Facet Random Model

In this study, a completely crossed, two-facet, random model design was developed to examine MFF latency and error scores at each of the four grade levels. Coefficients of generalizability were derived from this design. The two facets were: (1) items on the MFF (12 conditions corresponding to the 12 items) and (2) occasions (2 occasions, a test and a retest).

This analysis of variance (ANOVA) design yields estimates of seven components of variance. These components are given in Table 1. The observed score of a student on either error or latency for a two-facet design can be thought of as consisting of a grand mean plus seven components. The following equation shows the composition of an observed score:

$$X_{kmn} = \mu \qquad \text{(grand mean)}$$
$$+ \mu_n - \mu \qquad \text{(student effect)}$$
$$+ \mu_m - \mu \qquad \text{(item effect)}$$
$$+ \mu_k - \mu \qquad \text{(occasion effect)}$$
$$+ \mu_{nm} - \mu \qquad \text{(student x item interaction)}$$
$$+ \mu_{nk} - \mu \qquad \text{(student x occasion)}$$
$$+ \mu_{mk} - \mu \qquad \text{(item x occasion)}$$

$$+X_{kmn} - \mu_n - \mu_m - \mu_k - \mu_{nm} - \mu_{nk} \quad \text{(residual)}$$
$$-\mu_{mk} + \mu$$

Table 1 presents a balanced ANOVA design in which all effects are random. That is, the effects are assumed to be sampled from infinite universes. The letters k, m, and n are constants which represent the number of testing occasions, number of items, and number of students that are sampled in order to obtain estimates of variance components.

For the present study, the universe score for any student is the expected value of the average scores taken over all possible samples of admissible items and admissible occasions.

The expected observed score variance for students is given as: $$\sigma_x^2 = \sigma_s^2 + \frac{\sigma_{os}^2}{k} + \frac{\sigma_{si}^2}{m} + \frac{\sigma_e^2}{km}$$

where k is the number of testing occasions and m is the number of items in the G study. The sum of the last three terms of the equation is referred as $\sigma^2(\delta)$ by Cronbach et al. (1972). $\sigma^2(\Delta)$ would be the sum of $\sigma^2(\delta)$ with the variance estimates for items, occasions, and the interaction between items and occasions. In symbols:

$$\sigma^2(\delta) = \frac{\sigma_{os}^2}{k} + \frac{\sigma_{si}^2}{m} + \frac{\sigma_e^2}{km}$$

$$\sigma^2(\Delta) = \sigma^2(\delta) + \frac{\sigma_o^2}{k} + \frac{\sigma_i^2}{m} + \frac{\sigma_{oi}^2}{km}$$

Since the universe score is the expected value of the observed score taken over the universe of generalization, the variance for the universe score may be estimated by obtaining the limit of the observed score variance as both the number of items and occasions approaches infinity. To generalize over both items and occasions, the appropriate generalizability coefficient is:

$$\hat{\mathcal{E}}\rho^2(oI) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{os}^2}{k} + \frac{\sigma_{si}^2}{m} + \frac{\sigma_e^2}{km}} = \frac{\sigma_s^2}{\sigma_x^2}$$

Hence, $\hat{\mathcal{E}}\rho^2(oI)$ is approximately equal to the expected value of the correlation of two evaluations obtained from an independent random sample of k occasions and independent samples of m items. The sampling variation due to both occasions and items is included in the error variance term. Also included in this error term are the interactions between occasions and items with other unknown or confounded sources of variation. The σ_s^2 component is similar to the true score variance of classical test theory. In generalizability studies, it is used as an estimate of the universe score variance.

Method

Subjects

The subjects were 114 second, third, fourth, and fifth grade boys enrolled in a suburban Southern California school district. Subjects were white, middle-class to upper middle class, and were all enrolled full time in regular classrooms. There were 28 second grade boys (mean age 94.1 months, range 89-102), 26 third grade boys (mean age 106.8 months, range 100-116), 27 fourth grade boys (mean age 118.0 months, range 111-131), and 33 fifth grade boys (mean age 129.5, range 122-140). Four subjects, one at each grade level, moved before the study was completed. In addition, data for three subjects were discarded because examiner comments indicated that the test scores were not valid.

Procedure

Each child was individually administered Form F of the MFF using standard instructions and procedures (Kagan, 1965). Approximately three to six weeks later, each child was again administered the same test. Three female graduate students in school psychology who had been previously trained to administer the MFF served as examiners. Each boy was tested by a different examiner on

each occasion. The test was administered individually in a testing room at the school.

Results

Table 2 presents the means and standard deviations of MFF latency and error scores for the total sample and for each grade level on the two separate administrations of the test. From time 1 to time 2, there were decreases in mean error score at each grade level. In addition, error score decreases as grade level increases. With the exception of grade 5, there were decreases in mean latency score from time 1 to time 2. There was a small (1.36 seconds) increase in mean latency scores from time 1 to time 2 at grade 5.

Variance component estimates for MFF latency scores at each grade level are presented in Table 3. Variance component estimates for MFF error scores at each grade level are presented in Table 4. Following Cronbach et al. (1972), the negative variance components were set equal to zero. For MFF latency, the occasion and item x occasion interaction components were negative and set equal to zero. For MFF error scores, the student x item and item x occasion components at grade 2 and the item x occasion component at grade 4 were negative and set at zero.

The occasions component for MFF error at grade 5 was also set at zero.

The generalizability coefficients were obtained using a computer program developed by Erlich (1976). Table 5 presents the coefficients of generalizability for MFF latency and error for all combinations of one, two, and three occasions with 12 and 24 items. The first column presents the generalizability coefficients for 12 items and one occasion. This column is of particular interest because it provides an estimate of the dependability of the MFF as it has been commonly used in past research. The second column of Table 5 presents the generalizability coefficients for 12 items and two occasions which was the exact number of items and occasions which define the design of the current study. Columns three, four, five, and six present extrapolations which can serve as the basis for D studies involving the use of the MFF. These columns present estimates of the generalizability coefficients which would be obtained if the number of items and/or the number of testing occasions were increased.

The negative correlations between MFF latency and errors were moderate and significant at all grade levels,

ranging from -.52 at grade 4 to -.68 at grade 5.

Discussion

A central purpose of the present study was to present an application of generalizability theory for evaluation of the dependability of the MFF at four grade levels. There are several consistent patterns in the data. First, MFF latency is consistently a more dependable measure than MFF error at all grade levels. With a .70 generalizability coefficient as the criterion, adequate dependability is obtained for latency at grades three and five with only 12 items and one testing occasion. However, the MFF, in its most commonly used form, does not appear to provide a dependable measure of accuracy. The latency score, on the other hand, approached or exceeded the .70 criterion of dependability at the third, fourth, and fifth grades.

Secondly, both the number of items and the number of occasions make significant contributions to the generalizability of the MFF. It appears that the MFF, and in particular the error score, measures behavior that is susceptable to fluctuations over a short span of time. Increasing the number of occasions at which the MFF is administered thus results in a considerable increase in the generalizability of the scores. Moreover, the influence

of occasions appears to be relatively more potent than the influence of the number of items. In lengthening the test, the consistency of the MFF increases in terms of content sampling, but not in its stability over time.

A relatively dependable measure can be obtained by increasing the number of items to 24 and testing on tw separate occasions. Given the brevity of the 12 item MFF which is currently used, an increase in items and observations (occasions) would be feasible in many situations.

For the educational researcher who wishes to obtain a dependable measure of reflection-impulsiviey, a smaller number of items may be used if the number of testing occasions is increased. When retesting poses a problem, the researcher can still obtain a dependable measure by increasing the number of items given at one administration of the MFF, although the increase in dependability will not be as rapid as it would be if occasions were increased.

A third important finding of the present study pertains to the grade level differences in the dependability of the MFF. Kagan (1965) maintains that there is "a clear trend for decreasing errors and increasing response latencies with age" (p. 136). In addition, Kagan and Messer (1975) have argued that MFF errors and latency are more

highly correlated among older than younger children. Results of the present study suggest that, rather than reflecting a developmental change in cognitive style, the changes reported by Kagan (1965) and Kagan and Messer (1975) are function of the lower dependability of the MFF at younger ages. This can be seen by comparing the generalizability coefficients in Table 5 at grade two with those at the higher grade levels. Moreover, Table 2 shows no consistent pattern of increasing MFF latency with age, although there was a consistent pattern of decreasing error scores as grade level increases. These results are consistent with those reported by Egeland and Weinberg (1976).

As noted earlier, an important finding of the present investigation was the lower dependability of MFF error when compared to MFF latency. Egeland and Weinberg (1976), on the other hand, did not find a consistent pattern of lower test-retest reliability for MFF error. However, in contrast to a test-retest study which systematically evaluates only variance attributable to occasions, the present study considered the variance attributable to occasions as well as the variance attributable to items. Thus, the relatively lower dependability of the error scores obtained in the present study may reflect the

contribution of item variation to the coefficient of generalizability.

This possibility is particularly interesting in light of the factor analysis of MFF-like problems reported by Zelniker et al. (1977). Results of a principal components factor analysis of 40 MFF-like problems indicated that response latency is strongly determined by one factor and that response latency shows a high degree of internal consistency. In contrast, accuracy emerged as a complex factor associated with two principal factors. Internal consistency reliability was lower for accuracy than for latency and comparison with a dimension free index of internal consistency suggested that several dimensions exist in the error data. Thus, the lower dependability of MFF error and its relation to item variance in the present study may in part reflect an underlying multi-dimensional quality of MFF errors.

References

Ault, R. L., Mitchell, C., & Hartmann, D. P. Some methodological problems in reflection-impulsivity research. Child Development, 1976, 47, 227-231.

Bentler, P. M., & McLain, J. A multitrait-multimethod analysis of reflection-impulsivity. Child Development, 1976, 47, 218-226.

Block, J., Block, J. H., & Harrington, D. M. Some misgivings about the Matching Familiar Figures Test as a measure of reflection-impulsivity. Developmental Psychology, 1974, 10, 611-632.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Egeland, B., & Weinberg, R. A. The Matching Familiar Figures Test: A look at its psychometric credibility. Child Development, 1976, 47, 483-491.

Erlich, O. A study of the generalizability of measures of teacher behavior. Unpublished doctoral dissertation, University of California, Los Angeles, 1976.

Kagan, J. Impulsive and reflective children: Significance of conceptual tempo. In J. Krumboltz (Ed.), Learning and the educational process. Chicago: Rand McNally, 1965.

Kagan, J., & Kogan, N. Individual variation in cognitive processes. In P. Mussen (Ed.), Carmichael's manual of child psychology (Vol. 1). New York: Wiley, 1970.

Kagan, J., & Messer, S. B. A reply to "Some misgivings about the Matching Familiar Figures Test as a measure of reflection-impulsivity." Developmental Psychology, 1975, 11, 244-248.

Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. Information processing in the child: Significance of analytic and reflective attitudes. Psychological Monographs, 1964, 78, (1, Whole No. 578).

Zelniker, T., Bentler, P. M., & Renan, A. Speed versus accuracy as a measure of cognitive style: Internal consistency and factor analysis. Child Development, 1977, 48, 301-304.

Table 1 Summary of Random Effects ANOVA used for Each Grade Level and Each Dependent Variable.

Source of Variation	df	Expected Means-Squares
Students (S)	n-1	$km\sigma_s^2 + m\sigma_{os}^2 + k\sigma_{si}^2 + \sigma_e^2$
Items (I)	m-1	$kn\sigma_i^2 + n\sigma_{oi}^2 + k\sigma_{si}^2 + \sigma_e^2$
Occasions (O)	k-1	$nm\sigma_o^2 + m\sigma_{os}^2 + n\sigma_{oi}^2 + \sigma_e^2$
OS	(k-1)(n-1)	$n\sigma_{os}^2 + \sigma_e^2$
OI	(k-1)(m-1)	$n\sigma_{oi}^2 + \sigma_e^2$
SI	(n-1)(m-1)	$k\sigma_{si}^2 + \sigma_e^2$
OSI,error	(k-1)(n-1)(m-1)	σ_e^2

Table 2 Means and Standard Deviations of Total MFF Latency and Error Scores.

MFF Group	Time 1		Time 2	
	$\bar{X}$	s^2	$\bar{X}$	s^2
Total Sample (n=107)				
Latency	124.45	74.75	122.99	68.09
Error	11.65	5.77	9.36	5.17
Grade 2 (n=26)				
Latency	113.00	51.31	112.95	66.91
Error	14.81	5.24	12.04	5.01
Grade 3 (n=25)				
Latency	135.62	98.41	130.08	87.60
Error	13.16	5.84	9.52	5.16
Grade 4 (n=24)				
Latency	100.44	45.97	97.97	34.92
Error	12.04	4.57	9.75	3.37
Grade 5 (n=32)				
Latency	143.03	83.08	144.39	65.68
Error	7.63	4.75	6.75	5.32

Table 3 Estimates of Variance Components for MFF Latency Scores By Grade Level with the Design n x m x k.

MFF Latency

Source of Variance	Grade 2	Grade 3	Grade 4	Grade 5
Students (n)	13.3010	53.6655	7.32028	27.9920
Items (m)	2.2150	4.0685	1.41487	8.6110
Occasions (k)	0	0	0	0
Student-by-item (nm)	7.8797	2.43841	.930253	7.3039
Student-by- occasion (nk)	9.4289	3.77512	3.05313	7.7218
item-by- occasion (mk)	0	0	0	1.3478
residual (nmk, e)	15.5511	31.5485	13.47611	31.4652

Table 4 Estimates of Variance Components for MFF Error Scores By Grade Level with Design n x m x k.

	MFF Error			
Source of Variance	Grade 2	Grade 3	Grade 4	Grade 5
Students (n)	.034711	.118324	.031049	.093036
Items (m)	.141382	.096043	.151867	.057669
Occasions (k)	.021771	.040226	.015906	.000000
Student-by-item (nm)	.000000	.164982	.012379	.105341
Student-by-occasion (nk)	.045267	.079767	.003369	.043777
Item-by-occasion (mk)	.000000	.018951	.000000	.020477
Residual (nmk,e)	1.29112	.862591	.917808	.461216

Table 5 Generalizability coefficients, $\hat{\mathcal{E}}\rho^2$, for different number of items and occasions.

Grade	Type	I=12 O= 1	I=12 O= 2	I=12 O= 3	I=24 O= 1	I=24 O= 2	I=24 O= 3
2	Error	.1756	.2987	.3898	.2474	.3966	.4964
	Latency	.5390	.6891	.7588	.5606	.7132	.7833
3	Error	.5581	.6878	.7464	.6998	.8043	.8459
	Latency	.8902	.9402	.9577	.9118	.9533	.9680
4	Error	.2778	.4307	.5291	.4240	.5928	.6828
	Latency	.6317	.7717	.8332	.6667	.7987	.8550
5	Error	.5056	.6507	.7204	.5796	.7216	.7860
	Latency	.7188	.8288	.8731	.7497	.8530	.8944